Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 127–142

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Noise robust exemplar matching with alpha–beta divergence

Emre Yılmaz [a,*], Jort F. Gemmeke [b], Hugo Van hamme [a]

[a] Dept. ESAT-PSI, KU Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
[b] Audience Inc., 331 Fairchild Drive, 94043 Mountain View, CA, USA

Received 9 January 2015; received in revised form 6 October 2015; accepted 9 October 2015
Available online 17 October 2015

**Abstract**

The noise robust exemplar matching (N-REM) framework performs automatic speech recognition using exemplars, which are the labeled spectrographic representations of speech segments extracted from training data. By incorporating a sparse representations formulation, this technique remedies the inherent noise modeling problem of conventional exemplar matching-based automatic speech recognition systems. In this framework, noisy speech segments are approximated as a sparse linear combination of the exemplars of multiple lengths, each associated with a single speech unit such as words, half-words or phones. On account of the reconstruction error-based back end, the recognition accuracy highly depends on the congruence of the speech features and the divergence metric used to compare the speech segments with exemplars. In this work, we replace the conventional Kullback–Leibler divergence (KLD) with a generalized divergence family called the Alpha–Beta divergence with two parameters, $\alpha$ and $\beta$, in conjunction with mel-scaled magnitude spectral features. The proposed recognizer traverses the $(\alpha, \beta)$ plane depending on the amount of contamination to provide better separation of speech and noise sources. Moreover, we apply our recently proposed active noise exemplar selection (ANES) technique in a more realistic scenario where the target utterances are degraded by genuine room noise. Recognition experiments on the small vocabulary track of the 2nd CHiME Challenge and the AURORA-2 database have shown that the novel recognizer with the AB divergence and ANES outperforms the baseline system using the generalized KLD with tuned sparsity, especially at lower SNR levels.


*Keywords:* Automatic speech recognition; Noise robustness; Exemplar matching; Alpha–beta divergence; Reconstruction error

## 1. Introduction

Data-driven automatic speech recognition (ASR) techniques (De Wachter et al., 2003; Aradilla et al., 2005; Deselaers et al., 2007; Sundaram and Bellegarda, 2012; Sainath et al., 2012; Heigold et al., 2012; Sun et al., 2014) became popular in the last decade as a viable alternative after the long dominance of statistical acoustic modeling in the form of the Gaussian mixture models (GMM) in hidden Markov models (HMM) (Bourlard et al., 1996). Templates or exemplars are labeled speech segments of multiple lengths extracted from training data, each associated with a certain class, i.e. a speech unit such as phones, syllables or words. As they preserve the complete duration and trajectory information, exemplars are more immune to the inherent spectrotemporal variation of speech and its deteriorating effect on the ASR (Benzeghiba et al., 2007) compared to the conventional GMM–HMM- or deep neural networks (DNN)-based recognition systems. Moreover, it has been shown that using reasonably large exemplar sets overcomes the well-known generalization problem of the previous exemplar-based approaches (Seppi et al., 2010; Sun et al., 2011; Yılmaz et al., 2013a).

Exemplar matching-based recognition can be performed by evaluating the similarity of the exemplars with the

* Corresponding author. Tel.: +32 16 321828.
*E-mail addresses:* emre.yilmaz@esat.kuleuven.be (E. Yılmaz), jgemmeke@amadana.nl (J.F. Gemmeke), hugo.vanhamme@esat.kuleuven.be (H. Van hamme).

http://dx.doi.org/10.1016/j.specom.2015.10.004
0167-6393/

segments from the input speech with respect to a distance/divergence metric by applying dynamic time warping (Sakoe and Chiba, 1971; Ney and Ortmanns, 1999; De Wachter et al., 2007). In these applications, speech is represented using discriminatively trained features to ensure that the used distance/divergence metric mostly yields lower scores for the matching class compared to the other classes, resulting in increased recognition accuracies. The input speech segments can be simply classified as the label of the closest exemplar, or by a voting scheme on the set of K nearest neighbors (Golipour and O'Shaughnessy, 2009).

Exemplar-based sparse representations (SR) is an alternative data-driven ASR approach in which the spectrogram of input speech segments is modeled as a sparse linear combination of exemplars. SR-based techniques have been successfully used for speech enhancement (Gemmeke et al., 2011b), feature extraction (Sainath et al., 2010) and speech recognition (Kanevsky et al., 2010; Hurmalainen et al., 2011; Gemmeke et al., 2011a; Tan and Narayanan, 2012). These approaches model the acoustics using same-length exemplars labeled on the frame level and stored in a single overcomplete dictionary. The exemplar weights are obtained by solving a regularized convex optimization problem with a cost function consisting of the approximation quality with respect to a divergence and a term to induce sparse linear combinations using only a few exemplars. The choice of the divergence depends on the used speech features (how speech and noise sources are distributed in the high-dimensional feature space) to obtain reasonable sparse linear combinations. The non-negativity requirement of the SR formulation prevents the use of discriminatively trained features in this framework. The generalized Kullback–Leibler divergence (KLD) with the mel scaled magnitude spectral features has been successfully used in various applications in source separation, SR-based noise robust speech recognition and polyphonic music transcription (Virtanen, 2007; Smaragdis and Brown, 2003; Smaragdis, 2007; Raj et al., 2010; Tan and Narayanan, 2012). King et al. investigated the optimal parameter of the beta divergence as a cost function for non-negative matrix factorization-based speech separation and music interpolation in King et al. (2012).

This paper focuses on the divergence used by a recently proposed exemplar matching-based recognition approach, dubbed noise robust exemplar matching (N-REM) (Yılmaz et al., 2014a), which performs conventional exemplar matching in a SR formulation to be able to model noisy speech. Similar to the exemplar matching approaches, N-REM uses exemplars associated with a single speech unit such as phones, syllables, half-words or words. These exemplars are organized in separate dictionaries based on their duration (frame length) and class (associated speech unit). By applying a sliding window approach, the noisy speech segments are jointly approximated as a linear combination of the speech and noise exemplars using each dictionary. The recognizer adopts a reconstruction error based back-end, i.e. the recognition is performed by comparing the approximation quality for different classes quantified by a divergence measure and choosing the class sequence that minimizes the total reconstruction error.

The divergence plays an essential role in the recognition performance of N-REM on account of the reconstruction error based backend. The optimal divergence is expected to weight the individual reconstruction errors of each time–frequency cells in a way that the most informative cells contribute the most to the total reconstruction error. In this work, we use the Alpha–Beta (AB) divergence (Cichocki et al., 2011) in place of the generalized KLD to quantify the approximation error. The AB divergence is a family of divergences with two parameters, namely $\alpha$ and $\beta$. For different values of these parameters, the AB divergence connects various well-known distance/divergence measures such as the squared Euclidean distance, Hellinger distance, Itakura–Saito divergence and generalized KLD. The higher degree of freedom offered by the AB divergence has been shown to enable better robustness against noise and outliers (Cichocki et al., 2011).

The main contribution of this paper is a novel noise robust recognizer which traverses the ($\alpha$, $\beta$) plane based on the estimated SNR level to perform the most accurate separation of speech and noise sources. The recognition performance of the proposed system is investigated on the small vocabulary track of the 2nd CHiME Challenge (CHIME-2) and the AURORA-2 database. The initial ASR results at lower SNR levels ($-6$ dB and 0 dB of the CHIME-2 data) for numerous ($\alpha$, $\beta$) pairs are presented in Yılmaz et al. (2014b) and it has been shown that using AB divergence with an appropriate ($\alpha$, $\beta$) pair provides better recognition than the generalized KLD with tuned sparsity. In this work, we extend the recognition experiments to all SNR levels of both databases to have a better understanding of the novel system using the AB divergence. The baseline system which uses the generalized KLD as a dissimilarity measure is described in Yılmaz et al. (2014a). Secondly, an in-depth discussion on the impact of the divergence parameters on the recognition performance is provided by comparing the behavior of the generalized KLD and AB divergence for several ($\alpha$, $\beta$) pairs. Finally, we apply the adaptive noise modeling technique, active noise exemplar selection (ANES) (Yılmaz et al., 2014a), on the CHIME-2 data to investigate the recognition performance in case of genuine room noise. The rest of the paper is organized as follows. The N-REM using the AB divergence is described in Section 2. Section 3 discusses the evaluation setup and implementation details. Section 4 presents the recognition results and a discussion about the results is given in Section 5. Section 6 provides a general discussion and the concluding remarks.